Authorship ID at PAN'11

What -- Why -- How

Patrick Juola
Evaluating Variations in Language Laboratory
Duquesne University, Pittsburgh PA, USA
juola@mathcs.duq.edu

Authorship Identification

- … needs little definition among this group
- Differs subtly from plagiarism detection
  - Plagiarism : This part and THAT part differ
  - ID : This part is by THAT person
- But, yeah, still the same problem

Authorship Identification

- … needs little motivation among this group, either
  - School essays
  - Forged or disputed documents
  - Poison-pen letters (or Email)
  - Anonymous or corporate authorship

- Lots of reasons to study

… and lots of ways to do it

- Something of a "professional ad-hocracy"
- My own system (JGAAP) implements more than 1 million different approaches, most of which "work"
- … and none of which work perfectly

Hence, this track/lab

- NSF funded to create "community resources" to evaluate proposed methods
- NSF funded to create evaluation framework – i.e. on behalf of the NSF, welcome

This track : Email authorship

- Why one track? Possible better results from drilling down.
- Possible ability to re-use analysis; e.g. is one stemmer "better" than another?

- Why Email? Lots of data, and lots of importance.
  - If we had suggested a track on the Paston letters, who would have come?

Structure : 5 subtasks

- Closed class : 26 authors
- Closed class : 72 authors
- Open class : 26 authors
- Closed class : 72 authors
- Verification : 1 author at a time

Participants

- 31 registered groups /13 submissions8
- Scored by averaging precision, recall, and F score
- "Winners" :
  - Ludovic Tanguy (University of Toulouse & CNRS, France)
  - IoannisKourtis (University of the Aegean, Greece)
  - Mario Zechner (Know-Center, Austria)
  - Tim Snyder (Porfiau, Canada)

… but the real winner is the field

- … and everyone who participated
  - … or observed
    - … or is motivated to start looking further at this
- We hope to be back with an improved lab next year based on feedback here
- We hope to see you all back here with improved technology based on feedback here
- I look forward to seeing the papers!

Questions for next time

- New corpus, or extended corpus?
- Standardized markup?
- What languages/genres?
- What evaluation scheme?
- What other changes?

Dankuwel!